AUTOMATICWEB VIDEO CATEGORIZATION USING AUDIO-VISUAL INFORMATION AND HIERARCHICAL CLUSTERING RELEVANCE FEEDBACK

Bogdan Ionescu, Klaus Seyerlehner, Ionut Mironica, Constantin Vertan, Patrick Lambert

AUTOMATIC WEB VIDEO CATEGORIZATION USING AUDIO-VISUAL INFORMATION AND HIERARCHICAL CLUSTERING RELEVANCE FEEDBACK

B. Ionescu[†,*], *K. Seyerlehner*[§], *I. Mironică*[†], *C. Vertan*[†], *P. Lambert*[*]

[†]LAPI, University "Politehnica" of Bucharest, 061071, Romania,
{*bionescu, imironica, cvertan*}*@alpha.imag.pub.ro*
[§]DCP, Johannes Kepler University, A-4040 Austria,
klaus.seyerlehner@jku.at
[*]LISTIC, Polytech Annecy-Chambery, University of Savoie, 74944 France,
patrick.lambert@univ-savoie.fr

ABSTRACT

In this paper, we discuss and audio-visual approach to automatic web video categorization. We propose content descriptors which exploit audio, temporal, and color content. The power of our descriptors was validated both in the context of a classification system and as part of an information retrieval approach. For this purpose, we used a real-world scenario, comprising 26 video categories from the blip.tv media platform (up to 421 hours of video footage). Additionally, to bridge the descriptor semantic gap, we propose a new relevance feedback technique which is based on hierarchical clustering. Experiments demonstrated that retrieval performance can be increased significantly and becomes comparable to that of high level semantic textual descriptors.

Index Terms— audio-visual descriptors, video relevance feedback, web video genre classification.

1. INTRODUCTION

Automatic labeling of video footage according to genre is a common requirement in indexing large and heterogeneous collections of video material. In this paper we approach a specific *global* categorization problem: tagging videos according to *genre*. Feature extraction plays a key role in this process, as it provides the main representative power. Many sources of information have been tested [1]. *Text-based* information provides a higher semantic level of description than other sources and it was widely adopted by existing web platforms (e.g., YouTube, blip.tv). Text is obtained either from scene text (e.g., graphic text, sub-titles), from the transcripts of dialogues or from other external sources, for instance synopses, user tags, metadata. Common genre classification approaches include classic Bag-of-Words model and Term Frequency-Inverse Document Frequency (TF-IDF) approach [2]. *Audio-based* information can be derived either from time (e.g., Root Mean Square of signal energy - RMS, sub-band information, Zero-Crossing Rate - ZCR) of from the frequency domain (e.g., bandwidth, pitch, Mel-Frequency Cepstral Coefficients - MFCC). *Visual descriptors* exploit both static and dynamic aspects in the *spatial domain* or in the *compressed domain* (e.g., use of Discrete Cosine Transform coefficients). Spatial descriptors include *color information* (e.g., color histograms, predominant color, color entropy), *temporal structure* information (e.g., rhythm, average shot length, assessment of action), *object-based* information (e.g., occurrence of face and text regions, features such as Scale-Invariant Feature Transform - SIFT), and *motion* information (e.g., camera movement, object trajectory).

Although some sources of information provide better results than others in the context of video genre categorization [1], the most reliable approaches are however *multi-modal*. The approach in [3] combines spatial (face frames ratio, average brightness, and color entropy) and temporal (average shot length, cut percentage, average color difference and camera motion) descriptors and achieves a precision of up to 88.6% for the classification of movies, commercials, news, music, and sports (uses Support Vector Machines - SVM and Decision Trees). The approach in [4] uses visual-perceptual, structural, cognitive, and aural information with a parallel Neural Network based classifier. It achieves an accuracy rate up to 95% for the classification of football, cartoons, music, weather forecast, newscast, talk shows, and commercials. A generic approach to video categorization was discussed in [5]. Each video document is modeled with a Temporal Relation Matrix describing the relationship between video events, that is video segments with specific audio-visual patterns. The use of classification trees led to individual genre F_{score} ratios from 40% to 100% for news, soccer, TV series, documentary, TV games and movies.

We propose content descriptors which exploit both au-

[†]This work was supported under grant POSDRU/89/1.5/S/62557. We also acknowledge the 2011 Genre Tagging Task of the MediaEval Multimedia Benchmark for providing the test data set.

[§]This work was supported by the Austrian Science Fund L511-N15.

dio and visual (temporal and color) modalities. One of the novelties of our approach is in the way we compute the descriptors. The proposed *audio features* are block-level based, which compared to classic approaches have the advantage of capturing local temporal information [7]. *Temporal structure* information is based on the assessment of action perception at different levels. *Color information* is extracted globally, taking into account the temporal dimension. Compared to existing approaches which use mainly local or low-level descriptors the novelty of our approach is in the projection of color information onto a color naming system. This enables a higher perceptual level of description [6]. We performed the validation on a real-world categorization scenario and using a high diversity of genres, namely 26 video genres from blip.tv web media platform.

To bridge the inherent descriptor semantic gap resulted from the automatic nature of the annotation process, we investigate the potential use of relevance feedback techniques (RF). We propose a novel RF approach which is based on hierarchical clustering. Comparative experimental tests prove the potential of our approach which allows us to boost retrieval performance of the audio-visual descriptors close to that obtained with high-level semantic textual information.

The remainder of the paper is organized as follows: Section 2 presents the proposed video descriptors (audio, temporal, and color-based). Section 3 presents the proposed hierarchical clustering RF approach. Experimental validation is presented in Section 4, while Section 5 presents the conclusions and discusses future work.

2. VIDEO CONTENT DESCRIPTION

From the existing video modalities we exploit the *audio soundtrack*, *temporal structure*, and *color content*. Our selection is motivated by the specificity of these information sources with respect to video genre. Many genres have specific audio signatures, e.g., music clips contain rhythmic aspects, there is a higher prevalence of speech in news, in sports there is crowd noise, an so on. Considered visually, temporal structure and colors highlight specific genre contents; for instance, commercials and music clips have a high visual activity, music clips and movies use darker colors (as result of special effects), movies use cinematic principles, documentaries have reduced action content, sports have usually a predominant hue (e.g., green for soccer), in news broadcasting an anchorman is present (occurrence of faces).

Audio descriptors. The proposed set of audio descriptors are computed from sequences of consecutive spectral frames, called *blocks* and capture temporal properties (e.g., rhythmic aspects). Blocks are analyzed at a constant rate and are by default overlapping by 50% of their frames. Audio descriptors are computed by employing simple statistics separately over each dimension of the local block feature vectors (we employ mean, variance, or median statistics). After converting the soundtrack into a $22kHz$ mono signal, we compute [7]:

· *spectral pattern*: characterize the soundtrack's timbre via modeling those frequency components that are simultaneously active. Dynamic aspect of the signal are kept by sorting each frequency band of the block along the time axis.

· *delta spectral pattern*: captures the strength of onsets. To emphasize onsets, first we compute the difference between the original spectrum and a copy of the original spectrum delayed by 3 frames. Then, each frequency band is sorted along the time axis in a similar way as in the case of spectral pattern.

· *variance delta spectral pattern*: is basically an extension of the delta spectral pattern and captures the variation of the onset strength over time.

· *logarithmic fluctuation pattern*: captures the rhythmic aspects of the audio signal. In order to extract the amplitude modulations out of the temporal envelope in each band, periodicities are detected by computing the FFT along each frequency band of a block.

· *spectral contrast pattern*: roughly estimates the *"tone-ness"* of an audio track. For each frame within a block, we compute the difference between spectral peaks and valleys in several sub-bands. Then, the resulting spectral contrast values are sorted along the time axis in each frequency band.

· *correlation pattern*. To capture the temporal relation of loudness changes over different frequency bands, we use the correlation coefficients among all possible pairs of frequency bands within a block. The resulting correlation matrix forms the so-called correlation pattern.

Temporal structure descriptors capture information about the temporal structure:

· *rhythm*: capture the movie's tempo of visual change - we compute the relative number of shot changes occurring within a time interval of $T = 5s$, denoted ζ_T. Then, the rhythm is defined as the movie average shot change ratio, $\bar{v}_T = E\{\zeta_T\}$.

· *action*. We aim to highlight two opposite situations: video segments with a high action content (denoted "hot action") with $\zeta_T > 2.8$, and video segments with low action content with $\zeta_T < 0.7$. Thresholds were determined based on human observation of different action levels. This information is summarized with: hot-action ratio - $HA = T_{HA}/T_{video}$ and low-action - $LA = T_{LA}/T_{video}$, where T_X represents the total duration of all segments of type X.

· *gradual transition ratio*. Since high numbers of gradual transitions are generally related to a specific video editing style, we compute: $GT = (T_{dissolves} + T_{fade-in} + T_{fade-out})/T_{total}$ where T_X represents the total duration of all the gradual transitions of type X.

Color descriptors. Colors provide valuable information about the visual perception of the sequence. We project colors onto a color naming system, namely the Webmaster 216 color palette (colors are named according to hue, saturation, and intensity). We extend image-based color descriptors at

temporal level [6]:
· *global weighted color histogram*, h_{GW}, is computed as the weighted sum of each shot average color histogram. Weights are proportional to shot duration. The longer the shot the more important its contribution to the overall global histogram (values account for apparition percentage).
· *elementary color histogram*. We determine an elementary hue histogram by disregarding the saturation and intensity information: $h_E(c_e) = \sum h_{GW}(c)|_{Name(c_e) \subset Name(c)}$ where c_e is an elementary color (we use a total of 12 hues) and $Name()$ returns a color name from the dictionary.
· *color properties*. We define several color ratios. For instance, light color ratio, P_{light}, reflects the overall percentage of bright colors: $P_{light} = \sum h_{GW}(c)|_{W_L \subset Name(c)}$, where c is a color whose name contains one of the words defining brightness and $W_L \in \{$"light", "pale", "white"$\}$. Using the same reasoning and keywords specific to each property, we define dark color ratio (P_{dark}), hard saturated color ratio (P_{hard}), weak saturated color ratio (P_{weak}), warm color ratio (P_{warm}) and cold color ratio (P_{cold}). Additionally, we capture color richness with color variation, P_{var} (the number of different colors), and color diversity, P_{div} (the number of different hues).
· *color relationship*. We compute the percentage of similar perceptual colors (or adjacent, P_{adj}) and of opposite perceptual color pairs (or complementary, P_{compl}).

3. RELEVANCE FEEDBACK

To bridge the inherent descriptor semantic gap due to the automatic nature of the annotation process, we use relevance feedback (RF). RF takes advantage of the user expertise on the relevance of the results to compute a better representation of the information needed. We propose an RF approach which uses a hierarchical clustering mechanism (HC). HC has the advantage of producing a dendrogram representation which may be useful for displaying data and discovering data relationships. This mechanism allows us to select an optimal level from the dendrogram which provides a better separation of the relevant and non-relevant classes than the initial retrieval. The proposed RF scenario involves three steps: *retrieval*, *training*, and *updating*.

Retrieval. We provide an initial retrieval using a nearest-neighbor strategy. We return a ranked list of the N most similar videos to the query video using the Euclidean distance between features. This constitutes the initial RF window. Then, the user provides feedback by marking the relevant results, in our case movies from same genre category with the query.

Training. We initialize the clusters. Each cluster contains a single video from the initial RF window. We attempt to create two dendrograms, one for relevant videos and another for the non-relevant ones. We compute an initial cluster similarity matrix using the Euclidean distance between cluster centroids (which, compared to the use of min, max, and average distances, provided the best results). Then, we attempt to merge progressively the clusters from same relevance class using a minimum distance criterion. The process is repeated until the remaining number of clusters become relevant for the video categories within the retrieved window (we set this value to a quarter of N).

Updating. After finishing the training phase, we begin to classify the next videos as relevant or non-relevant with respect to the previous clusters. A given video is classified as relevant or not relevant if it is within the minimum centroid distance to a cluster in the relevant or non-relevant video dendrogram.

The main advantages of this RF approach are implementation simplicity and speed because it is computationally more efficient than other clustering techniques, such as SVM [8]. Further, unlike most RF algorithms (e.g., FRE [9] and Rocchio [10]), it does not modify the query or the similarity metric. The remaining retrieved videos are simply clustered according to class label.

4. EXPERIMENTAL RESULTS

The validation of the proposed content descriptors was carried out in the context of the MediaEval 2011 (Benchmarking Initiative for Multimedia Evaluation) Video Genre Tagging Task [2]. It addressed a real-world scenario - the automatic categorization of web videos from blip.tv. For tests we used a data set of 2375 sequences (around 421 hours) and 26 video genre categories (see Figure 1).

Classification scenario. In the first experiment we attempted to classify genres in terms of machine learning techniques. We used the Weka environment[1] and a cross-validation approach: data set is split into training and test sets and classification is repeated for random combinations between the two. The most accurate classification is obtained combining all audio-visual descriptors together (we use an early fusion approach). For reasons of brevity, we present only these results. Figure 2 shows the most significant results in terms of overall average (over all repetitions) correct classification.

The most accurate classification is obtained with SVM with linear kernel (namely 55% for 80% training, that is from 475 test sequences, 261 were correctly labeled), followed closely by Functional Trees (FT). The results are quite promising considering the high diversity of genres.

The most interesting results, however, were obtained at genre level. Due to the high semantic content, not all genres are to be accurately classified with audio-visual information. We sought to determine which categories are better suited for this approach. Figure 1 shows genre average F_{score}[2] for linear SVM and FT (numbers in brackets = number of test se-

[1] http://www.cs.waikato.ac.nz/ml/weka/
[2] http://www-nlp.stanford.edu/IR-book/

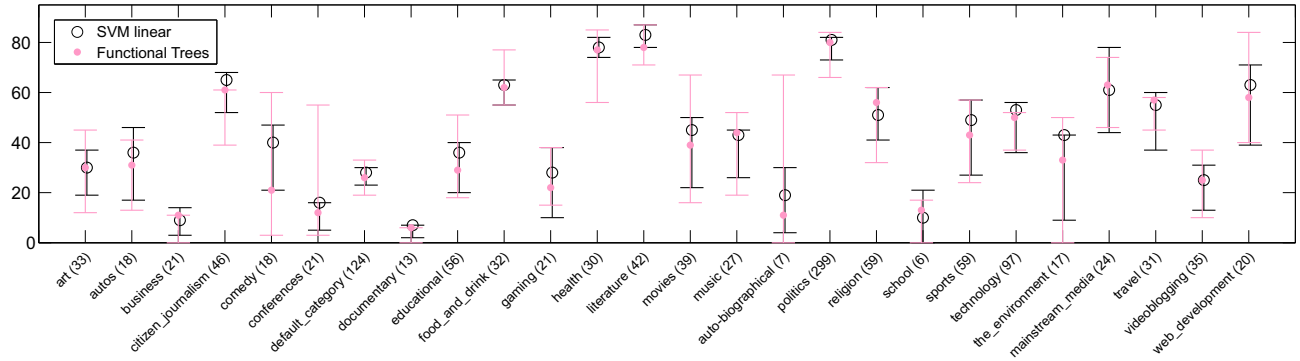

Fig. 1. Average F_{score} obtained with SVM and Functional Trees using all audio-visual descriptors (50% percentage split).

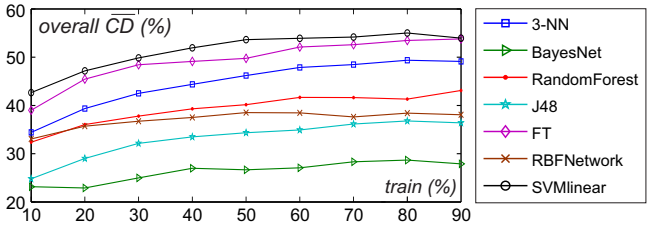

Fig. 2. Average correct classification for various machine learning techniques using all audio-visual descriptors.

quences/genre; vertical lines = min-max average F_{score} intervals for percentage split ranging from 10% to 90%).

The best performance is obtained for genres such as (we provide values for 50% percentage split and the highest): "literature" ($F_{score} = 83\%$, highest 87%) and "politics" ($F_{score} = 81\%$, highest 84%), followed by "health" ($F_{score} = 78\%$, highest 85%), "citizen journalism" ($F_{score} = 65\%$, highest 68%), "food and drink" ($F_{score} = 62\%$, highest 77%), "web development and sites" ($F_{score} = 63\%$, highest 84%), "mainstream media" ($F_{score} = 63\%$, highest 74%), and "travel" ($F_{score} = 57\%$, highest 60%). Due to the large variety of video materials, increasing the number of examples may result in overtraining and thus in reduced classification performance (e.g., in Figure 2, linear SVM for 90% training).

Retrieval scenario. We present the results obtained at MediaEval 2011 Video Genre Tagging Task [2]. The challenge of the competition was to develop a retrieval mechanism for all the 26 genres. Each participant was provided with a development set consisting of 247 sequences (unequally distributed with respect to genre). Participants were encouraged to build their own training set. We extended the data set to up to 648 sequences (using same source: blip.tv). The final retrieval task was performed only once on a test set consisting of 1727 sequences. We used the SVM linear classifier and all audio-visual descriptors. Retrieval results were provided using a binary ranking, where the maximum relevance of 1 is associated to the genre category to which the document was classified, while other genres have 0 relevance. Performance was assessed with Mean Average Precision (MAP; see trec_eval scoring tool[3]).

Table 1. MediaEval benchmark [2] (selection of results).

descriptors	modality	method	MAP	team
speech trans.	text	SVM	11.79%	LIA
speech trans., metadata, tags	text	Bag-of-Words + Terrier IR	11.15%	SINAI
speech trans.	text	Bag-of-Words	5.47%	SINAI
speech trans.	text	TF-IDF	6.21%	UAB
speech trans., metadata	text	TF-IDF + cosine dist.	9.34%	UAB
MFCC, zero cross., signal energy	audio	SVMs	0.1%	KIT
proposed	**audio**	**SVM**	**10.3%**	RAF
clustered SURF	visual	Bag-of-Vis.-Words+ SVM	9.43%	TUB
hist., moments, autocorr., co-occ., wavelet, edge hist.	visual	SVMs	0.35%	KIT
face statistics [4]	visual	SVMs	0.1%	KIT
shot statistics [4]	visual	SVMs	0.3%	KIT
proposed	**visual**	**SVM**	**3.84%**	**RAF**
color, texture, aural, cognitive, structural	audio, visual	SVMs	0.23%	KIT
proposed	**audio, visual**	**SVM**	**12.1%**	**RAF**

In Table 1 we compare our results with several other approaches using various modalities of the video, from textual information (e.g., provided speech transcripts, user tags, metadata) to audio-visual (Terrier Information Retrieval is to be found at `http://terrier.org/`). We achieved an overall MAP of up to 12.1% (see team RAF). These were the best results obtained using audio-visual information alone. Use of descriptors such as cognitive information (face statistics), temporal information (average shot duration, dis-

[3] `http://trec.nist.gov/trec_eval/`

tribution of shot lengths) [4], audio (MFCC, zero-crossing rate, signal energy), color (histograms, color moments, autocorrelogram - denoted autocorr.), and texture (co-occurrence - denoted co-occ., wavelet texture grid, edge histograms) with SVM resulted in a MAP of less than 1% (see team KIT), while clustered SURF features and SVM achieved a MAP of up to 9.4% (see team TUB). We achieved better performance even compared to some classic text-based approaches, such as the TF-IDF (MAP 9.34%, see team UAB) and the Bag-of-Words (MAP 11.15%, see team SINAI). It must be noted that the results presented in Table 1 cannot be definitive, as the classification approaches were not trained and set up strictly comparably (e.g., team KIT used up to 2514 sequences, most text-based approaches employed query expansion techniques). However, these results not only provide a (crude) performance ranking, but also illustrate the difficulty of this task. The most efficient retrieval approach remains the inclusion of textual information which conducted to an average MAP around 30% while the inclusion of information such as movie ID led to MAP up to 56% (the highest obtained).

Relevance feedback scenario. In the final experiment we sought to prove that, notwithstanding the superiority of text descriptors, audio-visual information also has great potential in classification tasks, but may benefit from help of relevance feedback. For tests, we used the entire data set: all 2375 sequences. Each sequence was represented by the proposed audio-visual descriptors. The user feedback was simulated automatically from the known class membership of each video (i.e., the genre labels). We use only one feedback session. Tests were conducted for various sizes of the user browsing window N, ranging from 20 to 50 sequences. Table 2 summarizes the overall retrieval MAP estimated as the area under the uninterpolated precision-recall curve.

Table 2. MAP obtained with Relevance Feedback

RF method	20 seq.	30 seq.	40 seq.	50 seq.
Rocchio [10]	46.8%	43.84%	42.05%	40.73%
FRE [9]	48.45%	45.27%	43.67%	42.12%
SVM [8]	47.73%	44.44%	42.17%	40.26%
proposed	$\mathbf{51.27\%}$	$\mathbf{46.79\%}$	$\mathbf{43.96\%}$	$\mathbf{41.84\%}$

For the proposed HCRF, the MAP ranges from 41.8% to 51.3%, which is an improvement over the other methods of at least a few percents. Also, it can be seen that relevance feedback proves to be a promising alternative for improving retrieval performance since it provides results close to those obtained with high-level textual descriptors (see Table 1).

5. CONCLUSIONS

We have addressed web video categorization and proposed a set of audio-visual descriptors. Experimental validation was carried out with a real-world scenario, using 26 video genres from the blip.tv web media platform. The proposed descriptors show potential in distinguishing for some specific genres. To reduce descriptor semantic gap, we have designed a relevance feedback approach which uses hierarchical clustering. It allowed boosting the performance close to the one obtained with higher-level semantic textual information. Future improvements will consist mainly of addressing sub-genre categorization and considering the constraints of very large scale approaches.

6. REFERENCES

[1] D. Brezeale, D.J. Cook, "Automatic Video Classification: A Survey of the Literature," IEEE TSMC, Part C: Applications and Reviews, 38(3), pp. 416-430, 2008.

[2] M. Larson, A. Rae, C.-H. Demarty, C. Kofler, F. Metze, R. Troncy, V. Mezaris, Gareth J.F. Jones (eds.), Working Notes Proceedings of the MediaEval 2011 Workshop at Interspeech 2011, vol. 807, Italy, September 1-2, 2011.

[3] X. Yuan, W. Lai, T. Mei, X.-S. Hua, X.-Q. Wu, S. Li, Automatic Video Genre Categorization using Hierarchical SVM," IEEE ICIP, pp. 2905-2908, 2006.

[4] M. Montagnuolo, A. Messina, "Parallel Neural Networks for Multimodal Video Genre Classification", Multimedia Tools and Applications, 41(1), pp. 125-159, 2009.

[5] Z. Al A. Ibrahim, I. Ferrane, P. Joly, "A Similarity-Based Approach for Audiovisual Document Classification Using Temporal Relation Analysis," EURASIP JIVP, doi : $10.1155/2011/537372$, 2011.

[6] B. Ionescu, D. Coquin, P. Lambert, V. Buzuloiu, "A Fuzzy Color-Based Approach for Understanding Animated Movies Content in the Indexing Task," EURASIP JIVP, doi:10.1155/2008/849625, 2008.

[7] K. Seyerlehner, M. Schedl, T. Pohle, P. Knees, "Using Block-Level Features for Genre Classification, Tag Classification and Music Similarity Estimation," Music Information Retrieval Evaluation eXchange, Netherlands, 2010.

[8] S. Liang, Z. Sun, "Sketch retrieval and relevance feedback with biased SVM classification," Pattern Recognition Letters, 29, pp. 1733-1741, 2008.

[9] Yong Rui, T. S. Huang, M. Ortega, M. Mehrotra, S. Beckman, "Relevance feedback: a power tool for interactive content-based image retrieval", IEEE Trans. on Circuits and Video Technology, 8(5), pp. 644-655, 1998.

[10] N. V. Nguyen, J.-M. Ogier, S. Tabbone, and A. Boucher, "Text Retrieval Relevance Feedback Techniques for Bag-of-Words Model in CBIR", Int. Conf. on Machine Learning and Pattern Recognition, 2009.